Achieve up to 10x Capacity Optimization with MapReduce Deduplication

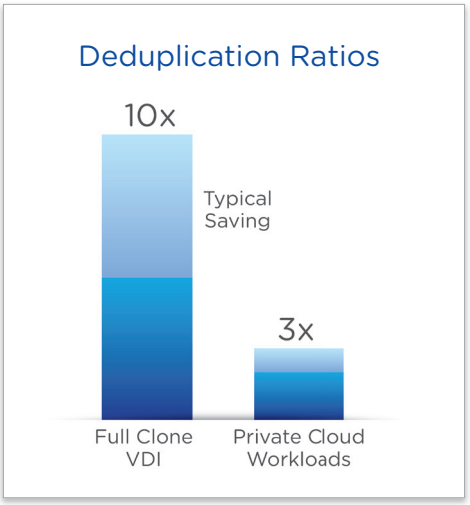
MapReduce Deduplication is a powerful post-process deduplication capability that can significantly boost the effective capacity of a cluster. Unlike deduplication in traditional storage arrays, MapReduce Deduplication is highly distributed and automatically runs on all nodes in the cluster. As nodes are added to the cluster, the deduplication functionality also linearly scales out.

Data is rapidly fingerprinted at the time it is written with a strong SHA-1 hash. The hashing process leverages performance acceleration offered by Intel processors for SHA-1. All nodes in the cluster run a background task scanning for fingerprints stored in their local cluster metadata. Through use of MapReduce redundant copies of data is eliminated in the storage capacity tier. MapReduce Deduplication, together with real-time performance deduplication already available in the system for the performance tier makes the Virtual Computing Platform ideal for persistent virtual desktop infrastructure (VDI) deployments and private cloud workloads.

Deduplication Results

Deduplication ratios vary based upon the type and nature of data.
The chart shows typical results expected with the Virtual Computing Platform.

All results assume that light to mediums workloads are being used for each use case. We also assume no 3rd party products\API's are being used to optimize space like linked clones or VAAI across a common operating system.



Use Case	Assumed VM Density per Node
VDI	75 - 125
Private Cloud Workloads*	30 - 60

*Private Cloud sizing taken from http://www.vmware.com/files/pdf/VMware-Service-Definition-Private-Cloud.pdf

Benefits of Scale-Out Deduplication

Unlike traditional storage solutions that can be bottlenecked at the controller level, Nutanix N-Controller model isn't impacted as the cluster grows in size delivering many benefits:

- No ripping and replacing controller heads as the dataset grows.
- Low overhead through only scanning metadata instead of the whole dataset.
- Provides the benefits of clones for environments where cloning wasn't an option and eliminates the complexity and support of maintaining 3rd party options.
- Intelligently focusing on OS and Application data that yields high results.
- Reduce the impact and time of Physical to Virtual & Virtual to virtual machine migrations.
- Hypervisor agnostic

Efficient Use of Resources

Contrary to traditional approaches which utilize background scans, requiring the data to be re-read, Nutanix performs the fingerprint in-line on ingest. Duplicate data in the capacity tier the data does not need to be scanned or re-read, essentially duplicate copies can be removed.

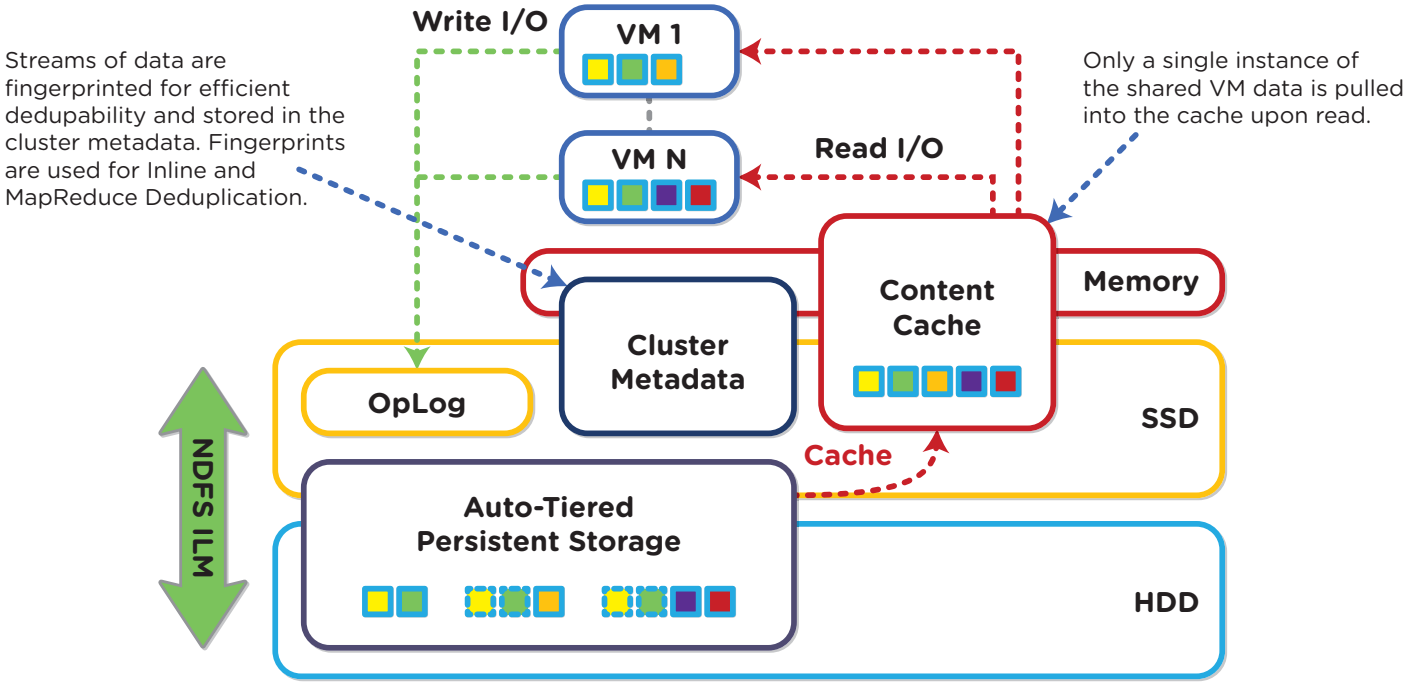

Write I/O
VM 1
VM N
Read I/O
Streams of data are fingerprinted for efficient dedupability and stored in the cluster metadata. Fingerprints are used for Inline and MapReduce Deduplication.
Only a single instance of the shared VM data is pulled into the cache upon read.
Content Cache
Memory
Cluster Metadata
OpLog
SSD
Cache
Auto-Tiered Persistent Storage
HDD
NDFS ILM

Figure 1 - One Nutanix node creating and saving fingerprints for MapReduce Dedupe

Frequently changing data even if it has the corresponding fingerprints computed will not be deduped until it's hasn't been written to for over an hour. This eliminates under undo stress on repeating work that will not have much value. Deduplication jobs also are throttled so running workloads are not affected.

MapReduce Deduplication is available in Pro and Ultimate editions and compliments inline deduplicaion from the Starter Edition. Simply by tuning MapReduce Dedupe on the cluster you can start to work where inline dedupe left off and achieve up to 10X more capacity.

Tel 855.NUTANIX | (855.688.2649)
Fax 408.916.4039
Email info@Nutanix.com
@nutanix

About Nutanix

Nutanix delivers web-scale IT infrastructure to medium and large enterprises with its software-driven Virtual Computing Platform, which natively converges compute and storage into a single solution to drive unprecedented simplicity of the datacenter. Customers can start with a few servers and scale to thousands, with fully predictable performance and economics. With a patented elastic data fabric and consumer-grade management, Nutanix is the blueprint for application-optimized infrastructure. Learn more at www.nutanix.com or follow us on Twitter @nutanix.

NUTANIX
Uncompromisingly Simple

www.nutanix.com

©2014 Nutanix, Inc. All Rights Reserved